Procena vrednosti fudbalera sa Transfermarkt-a primenom regresionih modela

Miloš Sirar, IN 3/2020, sirar.in3.2020@uns.ac.rs
Vasilije Zeković, IN 4/2020, zekovic.in4.2020@uns.ac.rs

I. Motivacija

Tema ovog projekta jeste procena vrednosti fudbalera iz pet najboljih liga na svetu, sa jednog od najrelevantnijih sajtova – Transfermarkt.

Fudbal je definitivno najpopularniji sport na planeti, koji se toliko razvijao kroz svoju istoriju, da danas to nije samo igra, već uveliko i biznis. To nisu samo fudbaleri, treneri i navijači, sada postoji posebna disciplina za svaki detalj i aspekt ovog sporta. Mnogo parametara se analizira, od same igre, fudbalera, taktike, terena... Jedna od vrlo važnih analiza jeste procena tržišne vrednosti fudbalera, koliko trenutno vredi svaki fudbaler u odnosu na mnoštvo različitih faktora (pozicija, klub, performanse...). Ova procena je od izuzetne važnosti da bi se predvidela kako trenutna vrednost fudbalera, tako i u budućnosti.

II. Opis početne baze podataka

Odabrana baza podataka se sastoji od 9 *csv* fajlova u kojima se nalaze podaci o igračima, nastupima i valuacijama igrača, klubovima, utakmicama, startnim postavama i takmičenjima. U ovakvom formatu je neupotrebljiva, pa je zato prvo analizirana da bi se izdvojili bitni parametri. Tokom analize, baze podataka su povezane i kreirana je jedinstvena baza podataka pod nazivom *all_data_new_3.csv*, koja sadrži 41 873 uzorka. Jedan uzorak u bazi podataka predstavlja jednu valuaciju vrednosti fudbalera koja sadrži informacije o fizičkim predispozicijama, statistici, pozicijama, klubu, takmičenju, kao i o vrednosti fudbalera u tom trenutku.

III. Analiza podataka

A. *Izdvajanje potrebnih obeležja i spajanje baza*

Iz baze podataka igrači, zadržana su obeležja: *državljanstvo*, *datum rođenja*, *pozicija*, *konkretna pozicija* i *visina*.

Baza podataka klubovi je služila za pronalazak klubova. Zaključeno je da su skoro sva obeležja irelevantna, osim obeležja *naziv kluba* i *ukupna vrednost kluba*. Ipak, odlučeno je da se za svaku godinu valuacije izračuna vrednost kluba pomoću vrednosti fudbalera u određenom periodu valuacije. Taj podatak je mnogo relevantniji nego da se uvek gleda samo vrednost kluba u 2023. godini.

Iz baze podataka valuacije igrača izdvojena su obeležja *vrednost fudbalera na datum valuacije* i *datum valuacije*.

Iz baze podataka takmičenje izdvojeni su samo nazivi i dva tipa takmičenja, domaće i evropsko takmičenje.

Nastupi koji su čuvani u bazi podataka sadrže veoma bitne informacije o statistici i performansama fudbalera i kao takvi zadržani su, ali u izmenjenom formatu. Umesto da se čuvaju podaci za svaki poseban nastup, oni su agregirani na način da se čuvaju statistike fudbalera u toku perioda između datuma valuacija, a pritom se čuva i informacija u kom klubu je fudbaler tada nastupao. Pored ovoga, statistike su podeljene na domaće takmičenje i evropsko takmičenje. Ova podela je dosta bitna, zbog toga što je učinak u evropskim takmičenjima značajniji i teže se dostiže taj nivo.

Nakon analize i pregleda obeležja i podataka iz preostale četiri baze podataka zaključeno je da su irelevantne.

B. *Nedostajuće vrednosti*

U ovako novokreiranoj bazi podataka otkriveno je postojanje nedostajućih vrednosti za obeležje *naziv kluba na datum valuacije*. Iako je ovo obeležje izbačeno, bilo je potrebno imati sve njegove vrednosti zbog kreiranja obeležja *vrednost kluba u godini valuacije*. Uočeno je postojanje null vrednosti za fudbalere koji nisu imali nastupe u periodu između dve valuacije. Velika većina ovih uzoraka čine rezervni golmani, koji retko imaju priliku da igraju utakmice. Nedostajuće vrednosti ovog obeležja su popunjene sa vrednostima za koji klub je fudbaler igrao u prvoj prethodnoj ili narednoj valuaciji.

C. *Analiza autlajera*

Sva numerička obeležja sadrže autlajere. Autlajeri svih obeležja predstavljaju stvarne ekstremne vrednosti, stoga ćemo ih zadržati.

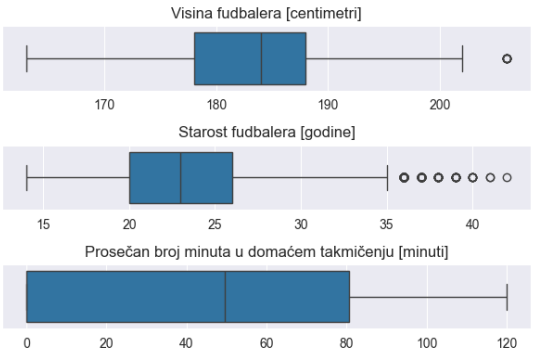
Obeležja koja su bila najviše zahvaćena ekstremnim vrednostima jesu *visina*, *starost*, *prosečan broj minuta u domaćem i evropskom takmičenju*. Na slici 1 može se videti kako izgleda osnovna raspodela tih obeležja.

Maksimalna vrednost za *visinu* iznosi 206 cm. Ovo se može objasniti time što golmani mogu da imaju dosta veću visinu od ostatka fudbalera, gde je prosek 183,17 cm.

Obeležje *starost* fudbalera se kreće u opsegu od 14 do 42 godine. Ove dve vrednosti su dosta izvan uobičajenih vrednosti za prosečne procene vrednosti fudbalera, zato je analizirano obeležje i došlo se do zaključka da su se stvarno proveravale vrednosti i za ovako mlade i stare fudbalere.

Kod preostala dva obeležja, *prosečnog broja minuta u domaćem i evropskom takmičenju* vidi se postojanje maksimalne vrednosti od 120 minuta. Ovo deluje čudno, s

obzirom da utakmica traje 90 minuta, ali se može objasniti time što postoje i produžeci od 30 minuta.



Slika 1. Box plot-ovi za prikaz autlajera

D. Analiza obeležja koje se predviđa – vrednost fudbalera na datum valuacije

Obeležje čija vrednost se predviđa je *vrednost fudbalera na datum valuacije* u milionima eura. Minimalna vrednost je 10 000 €, maksimalna vrednost 200 000 000€, prosek vrednosti je 9 646 029 €, a medijana 4 000 000 €.

Na slici 2 može se videti gustina raspodele vrednosti fudbalera, odnosno valuacija.

Slika 2. Gustina raspodele fudbalera u odnosu na njihovu vrednost, u stotinama milionima eura

Najveći broj valuacija, oko 87 % se nalazi u opsegu od „0“ do 20 miliona €. Zatim se primećuje pad broja valuacija čija je vrednost između 20 i 50 miliona €, koji iznosi oko 10%. Broj valuacija sa vrednošću od 50 do 200 miliona € iznosi oko 3% od ukupnog broja valuacija. Ovo je i očekivano, jer broj igrača koji su vrhunski i imaju veliku cenu je mnogo manji nego broj igrača koji imaju manju vrednost. Takođe i ti najbolji igrači su nekada imali manju vrednost dok nisu stigli do visokog nivoa.

E. Kategorička obeležja

Od 22 obeležja koja se nalaze u bazi podataka, 6 je kategoričko, dok je 16 numeričko. Kategorička obeležja su: *datum valuacije*, *državljanstvo*, *domaće takmičenje kluba na datum valuacije*, *pozicija*, *konkretna pozicija* i *jača noga*.

E.1 Pretvaranje kategoričkih obeležja u numerička – metod 1

Prvi način jeste stvaranjem kategorija koje predstavljaju zapravo brojevnu skalu.

Obeležje *datum valuacije* je prebačen u tertile od 2012. do 2023. godine, gde postoji ukupno 35 različitih vrednosti. Zbog velikog broja datuma valuacija koje je nemoguće pojedinčano grupisati svaka godina je podeljena na 3 perioda i svaka valuacija je raspoređena u odgovarajući tertil.

Kod obeležja *državljanstvo* se naišlo na veliki problem, pošto postoji čak 104 različite države. Ideja za rešavanja ovog problema jeste da se države podele u pet kategorija, gde je kategorija 5 najbolja, dok je kategorija 1 najgora kategorija. Na osnovu zvanične FIFA rang liste reprezentacija, države su raspoređene u odgovarajuću kategoriju. Bolje kategorije imaju manji broj država da bi do izražaja došli igrači koji igraju za bolje reprezentacije.

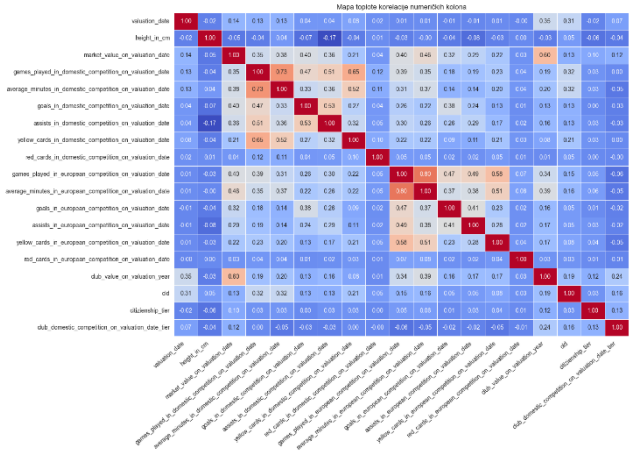
Obeležje *domaće takmičenje kluba na datum valuacije* ima sličan problem kao i državljanstvo, samo što ovde postoji ukupno 14 različitih liga. One su podeljene u tri kategorije, po jačini liga.

E.2 Pretvaranje kategoričkih obeležja u numerička – metod 2

Drugi način jeste kreiranje dummy varijabli za kategorička obeležja. Ovaj način je primenjen na preostala tri obeležja: *pozicija*, *konkretna pozicija* i *jača noga*. Zbog korišćenja dummy varijabli broj obeležja se povećao na 36. Iako je povećana dimenzionalnost ovaj način je jedini pravilan, zbog prirode obeležja.

F. Međusobna korelacija numeričkih obeležja

Na heatmap-i sa slike 3 može se videti korelacija između parova numeričkih obeležja. Primećena je jaka pozitivna korelacija između dva para obeležja: *broj odigranih utakmica u domaćem takmičenju na datum valuacije* sa *prosečnim brojem minuta u domaćem takmičenju na datum valuacije*, kao i za ista obeležja ali za evropsko takmičenje. Koeficijent korelacije je 0,73 kod domaćeg takmičenja, dok je kod evropskog takmičenja 0,8. Ovi koeficijenti su dosta logični, jer ukoliko fudbaler igra više utakmica, znači da će igrati i više minuta.



Slika 3. Prikaz korelacije između numeričkih obeležja pomoću heatmap-e

IV. METODE I MODELI MAŠINSKOG UČENJA

Nakon odabira, analize i promene baze podataka, došlo se do konačnog skupa podataka koji će se koristiti kao ulaz za različite regresione modele. Metode koje su odabrane: linearna regresija, kao osnovni model regresije, metoda *k* najbližih suseda, vrlo popularna zahvaljujući jednostavnosti i regresija na bazi vektora nosača (SVR), koja predstavlja kompleksniji model.

Skup podataka je podeljen na dva dela trening i test skup. Trening skup sadrži 90% podataka, dok preostalih 10% čini testni skup. Uzorkovanje je odrađeno nasumučno, postavljajući hiperparametar *shuffle* na true i *random_state* na vrednost konstante radi ponovljivosti rezultata. Kod trening skupa se koristi unakrsna validacija od pet podskupova (*n_splits* je 5) kako bi se preciznije utvrdile performanse usrednjavanjem procena svakog podskupa.

Za potrebe redukcije dimenzionalnosti korišćena je metoda razlaganja na glavne komponente (PCA), koja predstavlja najbolji izbor kod regresionih modela. Prilikom PCA transformacije odabrano je očuvanje varijanse od 99% (*n_components* je 0,99), koje donosi redukciju broja obeležja sa prvobitnih 35 na 25. Razlog za odabir varijanse od 99% je zadovoljavajuća redukcija obeležja i visoka očuvanost informacija iz originalne baze podataka.

Svaki model je dvostruko evaluiran – sa i bez PCA transformacije skupa podataka. Za evaluaciju modela i njihovo objektivno međusobno upoređivanje korišćena je srednja apsolutna greška (MAE). Ova metrika izračunava prosečnu apsolutnu razliku između predviđenih i stvarnih vrednosti, pružajući jasan uvid u preciznost modela. Pri proceni tržišne vrednosti fudbalera MAE metrika je vrlo ilustrativna, pošto nudi intuitivan i direktan pokazatelj tačnosti procene.

Dodatne metrike koje su korišćene pri proceni performansi regresora su: MSE predstavlja srednju kvadratnu grešku, slična je kao MAE samo što se razlike predviđenih i stvarnih vrednosti kvadriraju, umesto da se uzima njihova apsolutna vrednost. $RMSE$ je koren srednje kvadratne greške. R^2 skor je koeficijent determinacije, određuje meru objašnjene varijanse. $Adjusted\ R^2$ skor uzima u obzir koji prediktori koliko utiču na objašnjenu varijansu modela.

A. *Linearna regresija - LR*

Za linearnu regresiju korišćene su različite hipoteze, čiji rezultati se mogu videti u tabeli 1. Pored hipoteza prvog stepena, uvedene su i hipoteze višeg stepena zbog prisustva interakcija među obeležjima sa slike 3.

Tokom kreiranja hipoteza proveravani su i menjani hiperaparametri: *fit_intercept*, odnosno slobodan član hipoteze koji omogućava prolazak prave linearne regresije van koordinatnog početka, *include_bias* koji uvodi konstantu kao obeležje i *alpha* koji definiše visinu penalizacije koeficijenata pri regularizaciji.

Kroz proces testiranja različitih vrednosti hiperparametara, identifikovane su suboptimalne vrednosti za svaki model pojedinačno.

Vrsta hipoteze	$Avg.MAE$ (originalan skup)	$Avg.MAE$ (PCA skup)
I - Osnovna hipoteza	6 258 546	6 246 564
II - Osnovna hipoteza uz standardizaciju	6 258 546	-
III - Hipoteza sa interakcijama	$5{,}35 * 10^{14}$	5 239 639
IV - Hipoteza 2. stepena sa interakcijama	$5{,}18 * 10^{12}$	5 177 074
V - Hipoteza 3. stepena sa interakcijama	$2{,}27 * 10^{14}$	$1{,}12 * 10^{14}$
VI - Osnovna hipoteza uz Ridge regularizaciju	6 257 747	6 246 235
VII - Osnovna hipoteza uz Lasso regularizaciju	6 307 716	6 246 564
VIII - Hipoteza 2. stepena uz Ridge regularizaciju	**5 127 272**	5 241 824
IX - Hipoteza 3. stepena uz Ridge regularizaciju	5 255 197	6 239 676
X - Hipoteza 2. stepena uz Lasso regularizaciju	5 202 134	5 241 018
XI - Hipoteza 3. stepena uz Lasso regularizaciju	6 391 771	5 592 914

Tabela 1. Usrednjene vrednosti MAE procenjene unakrsnom validacijom nad modelima linearne regresije

U tabeli 1 kod hipoteza sa interakcijama vidi se problem sa određenim koeficijentima i stoga se dobijaju izuzetno velike vrednosti MAE. Najveći koeficijent kod modela III (bez PCA) iznosi $\approx 1{,}0065 * 10^{18}$, dok je kod modela VIII (bez PCA) $\approx 8{,}302 * 10^{6}$. Rezultati koeficijenata bi se mogli objasniti visokom multikolinearnošću među obeležjima, odnosno postojanjem različitih kombinacija obeležja ili samih obeležja koja redundantno reprezentuju isto, što rezultira koeficijentima koji izražavaju slične ili iste odnose. Model teže razlikuje doprinose pojedinačnih obeležja ciljnoj varijabli. Razlog tome može biti izbacivanje obeležja *player id* iz modela. Nakon toga, model je ostvarivao lošiji MAE, ranije je to bilo oko 3 miliona € u najboljem slučaju, dok je sada reč o 5 miliona €. Pored toga, ranije nije postojao problem da MAE bude za nekoliko redova veličina veći, uvek je bio oko 10^6. Razlog zbog kog se ovo dešava leži u hvatanju obrazaca i grupisanju svakog uzorka - valuacije u okviru nekog konkretnog igrača od strane modela. Pristup sa eliminisanjem *player id-a* je svakako pravilniji, i model iako poseduje veću grešku predstavlja realniji prikaz i bolje će generalizovati na neviđenim podacima. Biće navedena dva obrazloženja zašto je loše zadržati *player id*. Ako bi se model koristio u praksi potrebno je da se zna *player id* za novog igrača, ali kako ga odrediti? Zatim ako postoji više uzoraka koji su vezani za određenog igrača i deo uzoraka bude smešten u trening a deo u test skupu, nedvosmisleno će doći do curenja informacija i model će se natprilagoditi podacima. Bitno je spomenuti da je vrlo verovatno došlo do natprilagođenja podacima od strane modela sa hipotezama višeg stepena i interakcijama.

Regularizacijom je uspela penalizacija koeficijenata koji ne doprinose modelu, kao i umanjivanje onih koji preuveličavaju svoj značaj.

B. *Metoda k najbližih suseda - kNN*

Sledeća metoda koja se koristi jeste je *k* nearest neighbors regressor. Primenjeni su kNN modeli sa i bez standardizacije, odnosno sa i bez PCA transformacije. Standardizacijom kod kNN rešava se problem da ukoliko postoje obeležja kod kojih je raspon na osi izuzetan kao i obeležja kod kojih je raspon na osi minimalan, uz upotrebu metrike bazirane na rastojanju moguće je posledično

narušiti performanse regresora.

Obavljena je standardna procena hiperparametara unakrsnom validacijom, gde je korišćena Hemingova, Euklidska, Menhetn, Minkovski i Čebišljeva metrika (hiperparametar metric).

Isprobane su dve verzije za ponderisanje udaljenosti suseda gde se koristi hiperparametar *weights*. Prva verzija ima vrednost *distance* – sa ponderisanjem, dok druga verzija ima vrednost *uniform* – bez ponderisanja. Metoda je testirana za k iz skupa vrednosti $\{1, 2, 3, 4, 5\}$ najbližih suseda koje treba uzeti pri računanju predviđene vrednosti.

Hemingova metrika se može pokazati dobro zbog upoređivanja vrednosti obeležja – fudbaleri koji imaju većinu istih osobina verovatno će vredeti isto. Mana je što zbog velikih opsega mogućih vrednosti statistika fudbalera retko bude većih poklapanja dimenzija. Euklidska ima velike šanse da se pokaže najbolje, jer fudbaleri koji su izuzetno bliskih vrednosti obeležja nalaze se na maloj udaljenosti u prostoru obeležja. Sličan način zaključivanja vredi i za Menhetn, Čebišljevu i Minkovski metriku, s tim da se kod Minkovski metrike parametar *p* bira takav da bude različit od 1, 2 i beskonačno, kako se ne bi ponovile prethodne. Za *p* je odabrana vrednost 5.

	XII – kNN regresor (originala n skup)	XIII – kNN regresor (PCA skup)	XIV – kNN regresor uz standardizaciju (originalan skup)
k	4	5	5
metric	Menhetn	Euklidska	Menhetn
weights	distance	distance	distance
$Avg.\ MAE$	4 924 089	5 030 047	**4 653 685**

Tabela 2. Usrednjene vrednosti MAE procenjene unakrsnom validacijom nad kNN modelima

Iz tabele 2 se može uočiti da argumenti koji su ranije navedeni opravdavaju dobijene rezultate. Metrike poput Euklidske i Menhetn su od najvećeg doprinosa, zbog baziranja u njihovoj osnovi na geometrijskom rastojanju, odnosno bliskosti uzoraka u prostoru obeležja. Veće vrednovanje bližih suseda za ovakav problem prirodno daje bolje rezultate. Standardizacija je pokazala snagu u svođenju vrednosti obeležja na slične skale.

C. *Support vector regression - SVR*

SVR je prilično dobar izbor u pogledu regresionih modela. Biće navedeni neki od argumenata. Omogućava toleranciju po pitanju broja uzoraka koji se nalaze sa unutrašnje strane u odnosu na marginu (dobra strana margine). U stanju je da koristi različite kernel funkcije, koje obezbeđuju adekvatno formiranje modela, a koji je dalje u stanju da najbolje opiše podatke. Pomoću parametra *C* moguće je kontrolisati vrednost penalizovanja uzoraka sa pogrešne strane margine. Samim tim se direktno utiče na bias-variance trade off. Sa povećanjem *C*, model postaje više fleksibilan, varijansa raste, a pristrasnost opada, dok je sa smanjenjem *C* obrnuto. Radijalni kernel je nezaobilazan za većinu regresionih problema, gde je ključni parametar γ koji će odrediti širinu kernela i koje uzorke treba uzeti u obzir. Hiperparametar *degree* omogućava polinomijalnom kernelu da složenije fit-uje podatke, dok *coef0* obezbeđuje da slobodan član ne bude isključivo 0.

Budući da je cilj evaluacija performansi regresora nad što više kombinacija vrednosti hiperparametara to predstavlja računarski prilično zahtevan proces. Jedna opcija je da se smanji broj vrednosti hiperparametara, što je prilično loše, jer postoji mogućnost izbacivanja značajne vrednosti, dok je druga opcija da se skup uzoraka smanji nasumičnim odabiranjem i zatim izvrši unakrsna validacija. Odabrana je druga ideja. Skup je prethodno standardizovan iz prostog razloga kako bi se opsezi vrednosti sveli na približno istu skalu i time poboljšala moć modela.

	XV - SVR (originalan skup)	XVI - SVR (PCA skup)
C	10	10
coef0	2	2
degree	8	8
γ	scale	Auto
kernel	poly	Poly
$Avg.\ MAE$	**6 126 518**	6 135 984

Tabela 3. Usrednjene vrednosti MAE procenjene unakrsnom validacijom nad SVR modelima

Pretpostavka je da je polinomijalni kernel stepena 8 izabran zbog postojanja nelinearnih odnosa između obeležja. Parametar *C* je nešto veći što doprinosi tačnosti rezultata. Visok degree i *C* mogu biti problematični zbog natprilagođenosti. Parametar γ je procenjen adaptivno u skladu sa brojem obeležja.

V. ZAKLJUČAK

Do sada su bile prikazivane samo MAE metrike. Na kraju je za svaku od tri regresione metode odabran model sa najboljim hiperparametrima. Nakon toga sledi obuka nad čitavim trening skupom i evaluacija modela pomoću i ostalih metrika. Dobijeni rezultati su prikazani u tabeli 4.

	VIII - Hipoteza 2. stepena uz Ridge regularizaciju (originalan skup)	XIV – kNN regresor uz standardizaciju (originalan skup)	XV – SVR (originalan skup)
MSE	$\mathbf{6,484 * 10^{14}}$	$7{,}294 * 10^{14}$	$11 * 10^{14}$
MAE	4 995 183	**4 555 908**	5 417 564
$RMSE$	**8 052 317**	8 540 372	10 488 854
R^2	**0,667**	0,625	0,435
$Adj.\ R^2$	**0,661**	0,625	0,434

Tabela 4. Procena performansi tri najbolja regresiona modela

Model $VIII$ se izdvaja kao najpouzadniji, kako pokazuje R^2 i $Adjusted\ R^2$ skor, uprkos najvećem MAE modela $XIII$. Zaključak je da je bolje koristiti model $VIII$, jer će se bolje ponašati u različitim scenarijima.